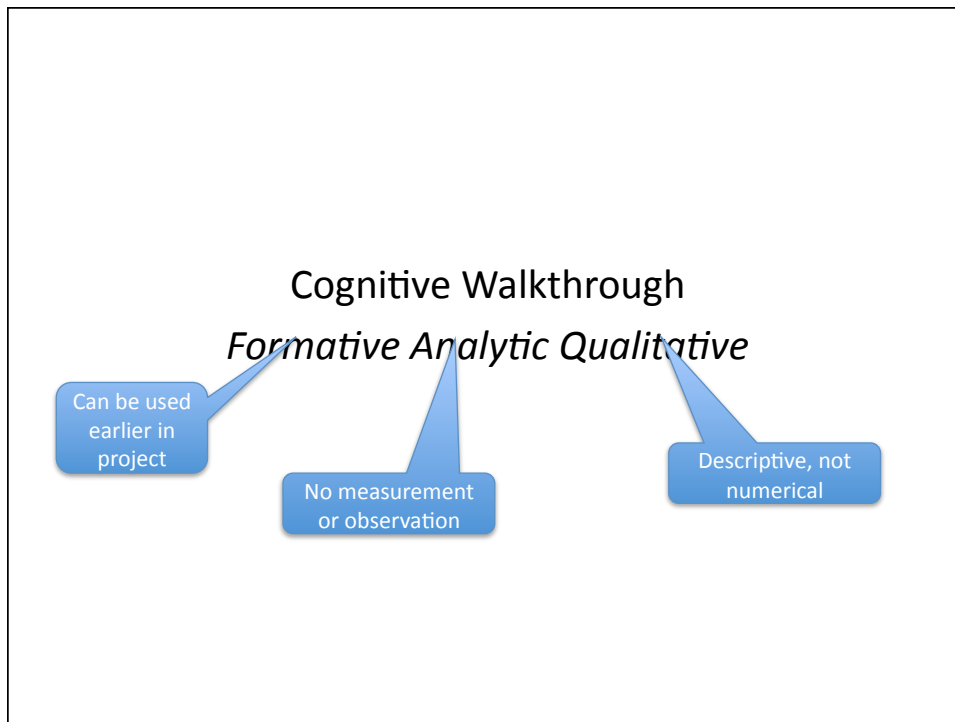


## Human-Computer Interaction

### Lecture 8: Usability evaluation methods

#### Different kinds of system evaluation/research

- Analytic/Empirical
  - ‘Analytic’ means reasoning and working by *analysis*
  - ‘Empirical’ means making *observations* or *measurements*
- Formative/Summative
  - Formative research (earlier in a project) evaluates & refines *ideas*
  - Summative research (later in a project) tests & evaluates *systems*
- Qualitative/Quantitative
  - Qualitative data involves *words* (or pictures), and can provide broad / detailed information about a small number of users and their context.
  - Quantitative data involves *numbers*, and can be used to compare data from larger numbers of users, or measure some specific aspect of their behaviour.



### From cognitive theory of exploratory learning

- User sets a **goal** to be accomplished, in terms of the expected system capabilities.
- User searches interface for currently available **actions**.
- User **selects** the action that seems likely to make progress toward the goal.
- User **performs** the selected action and **evaluates** the feedback given by the system, looking for evidence that progress has been made.
  - The user learns what to do in future by observing what the system does

## Evaluation procedure

- Manually simulate an (*imaginary*) user carrying out the stages of the model.
  - relies on knowing enough about this person to anticipate their prior knowledge / mental model.
- Evaluators move through task, telling a *story* about why user would choose each action.
- Evaluate the story according to:
  - user's current *goal*.
  - *accessibility* of correct control.
  - quality of *match* between label and goal.
  - *feedback* after the action.

## GOMS

*Formative Analytic Quantitative*

Can be used  
with partial  
implementation

No measurement  
or observation

Provides  
numerical data

## GOMS: Goals, Operators, Methods, Selection

- Goals: what is the user trying to do?
- Operators: what actions must they take?
  - **H**ome hands on keyboard or mouse
  - **K**ey press & release (tapping keyboard or mouse button)
  - **P**oint using mouse/lightpen etc
- Methods: what have they learned in the past?
- Selection: how will they choose what to do?
  - **M**ental preparation

## Interviews and Ethnographic Studies

*Formative Empirical Qualitative*

Can be used  
from start of  
project

Involves  
observation

Descriptive, not  
numerical

## Structured interviews

- Additional to requirements definition meetings.
- Encourage participation from a range of users.
- *Structured* in order to:
  - collect data into common framework
  - ensure all important aspects covered
- Newman & Lamming's proposed structure:
  - *activities, methods* and *connections*
  - *measures, exceptions* and *domain knowledge*
- Semi-structured interviews:
  - Ask further questions to probe topics of interest

## Observational task analysis

- Less intrusive than interviews
- Potentially more objective
- Inspired huge debate between cognitive and sociological views of HCI: see Lucy Suchman
- Harder work:
  - transcription from video protocol
    - relative duration of sub-tasks
    - transitions between sub-tasks
    - interruptions of tasks
  - alternatively, transcription from audio recording

## Ethnographic field studies

- Field observation to understand users and context
- Division of labour and its coordination
- Plans and procedures
  - When do they succeed and fail?
- Where paperwork meets computer work
- Local knowledge and everyday skills
- Spatial and temporal organisation
- Organisational memory
  - How do people learn to do their work?
  - Do formal methods match reality?
- See Beyer & Holtzblatt, *Contextual Design*

## Controlled Experiments

*Summative Empirical Quantitative*

Suitable for  
end of  
project

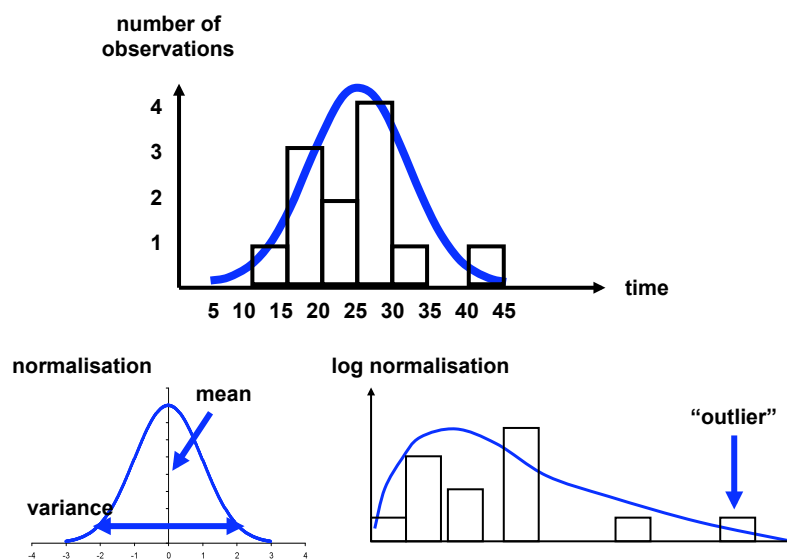
Involves  
measurements

Provides  
numerical data

## Controlled experiments

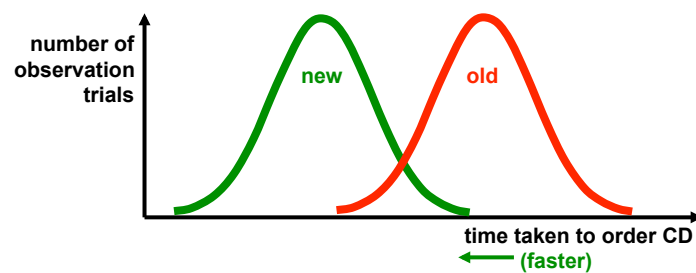
- Based on a number of observations:
  - How *long* did Fred take to order a CD from Amazon?
  - How many *errors* did he make?
- But every observation is different.
- So we compare averages:
  - over a number of trials
  - over a range of people (experimental participants)
- Results often have a normal distribution

## (statistics: histograms & distributions)



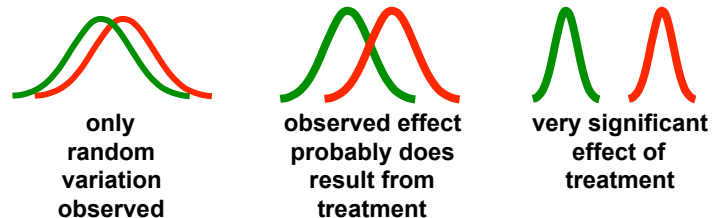
## Experimental treatments

- A *treatment* is some modification that we expect to have an effect on usability:
  - How long does Fred take to order a CD using this great new interface, compared to the crummy old one?
  - Expected answer: *usually* faster, but not *always*



## Hypothesis testing

- *Null hypothesis*:
  - What is the probability that this amount of difference in means could be random variation between samples?
  - Hopefully very low ( $p < 0.01$ , or 1%)
  - Use a statistical *significance test*, such as the *t-test*.



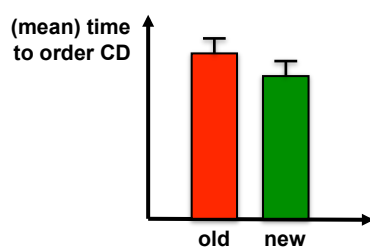


## Sources of variation

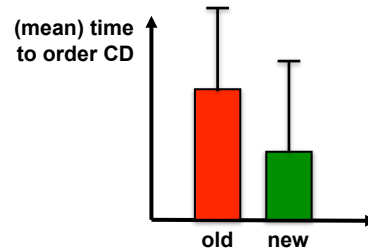
- People differ, so quantitative approaches to HCI must be statistical.
- We must distinguish sources of variation:
  - The effect of the treatment - what we want to measure.
  - Individual differences between subjects (e.g. IQ).
  - Distractions during the trial (e.g. sneezing).
  - Motivation of the subject (e.g. Mondays).
  - Accidental intervention by experimenter (e.g. hints).
  - Other random factors.
- Good experimental design and analysis isolates these.

## Effect size – means and error bars

- Difference of two means may be statistically significant (if sample has low variance), without being very interesting.
  - But mean differences must *always* be reported with a confidence interval, or plotted with 'error bars'



Experiment A: 'significant' but boring



Experiment B: interesting, but treat with caution

## Problems with controlled experiments

- Huge variation between people (~200%)
- Mistakes mean huge variation in accuracy (~1000%)
- Improvements are often small (~20%)
- ... or even negative (because new & unfamiliar)
- Most people give up using a new product at learning time anyway, so quantitative measures of 'expert' speed and accuracy performance may not be of great commercial interest
  - We don't care if it's slow, so long as users like it
  - (and user's perception of speed is inaccurate anyway)

## Surveys and Questionnaires

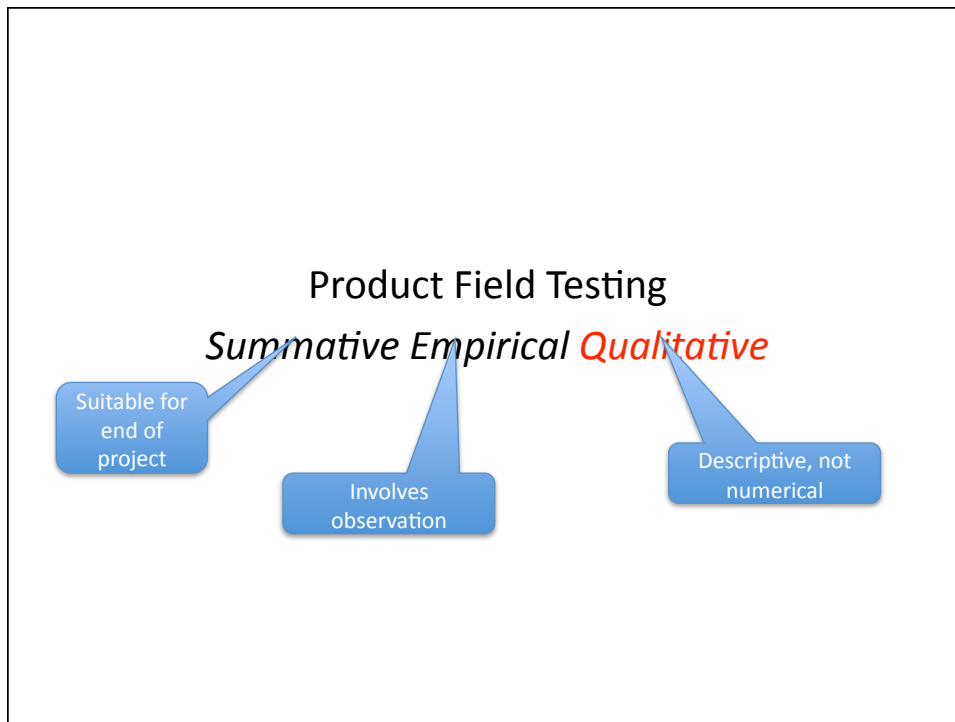
*Self-report measures*

## Surveys and questionnaires

- Standardised *psychometric instruments* can be used
  - To evaluate mental states such as fatigue, stress, confusion
  - To assess individual differences (IQ, introversion ...)
- Alternatively, questionnaires can be used to collect *subjective* or *self-report* evaluation from users
  - as in market research / opinion polls
  - ‘I like this system’ (and my friend who made it)
  - ‘I found it intuitive’ (and I like my friend)
- This kind of data can be of limited value
  - Can be biased, and self-report is often inaccurate anyway
  - It’s hard to design questionnaires to avoid these problems

## Questionnaire design

- *Open* questions ...
  - Capture richer qualitative information
  - But require a *coding frame* to structure & compare data
- *Closed* questions ...
  - Yes/No or *Likert* scale (opinion from 1 to 5)
  - Quantitative data easier to compare, but limited insight
- Collecting survey data via interviews gives more insight but questionnaires are faster
  - Can collect data from a larger sample
  - Remember to test questionnaires with a pilot study, as it’s easier to get them wrong than with interviews



## Product field testing

- Brings advantages of task analysis/ethnography to assessment and testing phases of product cycle.
- Case study: Intuit Inc.'s Quicken product
  - originally based on interviews and observation
  - follow-me-home programme after product release:
    - random selection of shrink-wrap buyers;
    - observation while reading manuals, installing, using.
  - Quicken success was attributed to the programme:
    - survived predatory competition from Microsoft Money
    - later valued at \$15 billion.

## *Non-Evaluation*

### Bad evaluation techniques

- Purely *affective* reports: 20 subjects answered the question “Do you like this nice new user interface more than that ugly old one?”
  - Apparently empirical/quantitative
- No testing at all: “It was decided that more colours should be used in order to increase usability.”
  - Apparently formative/analytic
- Introspective reports made by a single subject (often the programmer or project manager): “I find it far more intuitive to do it this way, and the users will too.”
  - Apparently analytic/qualitative

## Evaluation in Part II projects

### Summary of analytic options (analysing your design)

- Cognitive Walkthrough
  - Normally used in formative contexts – if you do have a working system, then why aren't you observing a real user (far more informative than simulating/imagining one)?
  - But Cognitive Walkthrough can be a valuable time-saving precaution before user studies start, to fix blatant usability bugs
- GOMS
  - unlikely you'll have alternative detailed UI designs in advance
  - If you have a working system, a controlled observation is superior
- Cognitive Dimensions
  - better suited to less structured tasks than CW & GOMS, which rely on predefined user goal & task structure

## Summary of empirical options (collecting data)

- Interviews/ethnography
  - could be useful in formative/preparation phase
- Think-aloud / Wizard of Oz
  - valuable for both paper prototypes and working systems
  - can uncover usability bugs if analysed rigorously
- Controlled experiments
  - appears more 'scientific', but only:
    - If you can measure the important attributes in a meaningful way
    - If you test significance and report confidence interval of observed means
- Questionnaires
  - be clear what you are measuring – is self-report accurate?
- Field Testing
  - controlled release (and data collection?) may be possible
- See human participants guidance for empirical methods

## Evaluation options for non-interactive systems

- Should your evaluation be analytic or empirical?
  - How consistent / well-structured is your analytic framework?
  - What are you measuring & why? Are the measurements compatible with your claims (*validity*)?
- Should your evaluation be formative or summative in nature?
  - If formative – couldn't you finish your project?
  - If summative – are the criteria internal or external?
- Is your data quantitative or qualitative?
  - Descriptive aspects of the system, or engineering performance data?
  - If qualitative, how will you establish objectivity (i.e. that this is not simply your own opinion)?